

Information Structure in African Languages: Corpora and Tools

Christian Chiarcos*, Ines Fiedler**, Mira Grubic*, Andreas Haida**, Katharina Hartmann**, Julia Ritz*, Anne Schwarz**, Amir Zeldes**, Malte Zimmermann*

* Universität Potsdam
Potsdam, Germany
{chiarcos|grubic|
julia|malte}@
ling.uni-potsdam.de

** Humboldt-Universität zu Berlin
Berlin, Germany
{ines.fiedler|andreas.haida|
k.hartmann|anne.schwarz|
amir.zeldes}@rz.hu-berlin.de

Abstract

In this paper, we describe tools and resources for the study of African languages developed at the Collaborative Research Centre “Information Structure”. These include deeply annotated data collections of 25 subsaharan languages that are described together with their annotation scheme, and further, the corpus tool ANNIS that provides a unified access to a broad variety of annotations created with a range of different tools. With the application of ANNIS to several African data collections, we illustrate its suitability for the purpose of language documentation, distributed access and the creation of data archives.

1 Information Structure

The Collaborative Research Centre (CRC) “Information structure: the linguistic means for structuring utterances, sentences and texts” brings together scientists from different fields of linguistics and neighbouring disciplines from the University of Potsdam and the Humboldt-University Berlin. Our research comprises the use and advancement of corpus technologies for complex linguistic annotations, such as the annotation of information structure (IS). We define IS as the structuring of linguistic information in order to optimize information transfer within discourse: information needs to be prepared (“packaged”) in different ways depending on the goals a speaker pursues within discourse.

Fundamental concepts of IS include the concepts ‘topic’, ‘focus’, ‘background’ and ‘information status’. Broadly speaking, the topic is the entity a specific sentence is construed *about*, focus represents the *new* or *newsworthy* information a sentence conveys, background is that part of the sentence that is *familiar* to the

hearer, and information status refers to different *degrees of familiarity* of an entity.

Languages differ wrt. the means of realization of IS, due to language-specific properties (e.g., lexical tone). This makes a typological comparison of traditionally less-studied languages to existing theories, mostly on European languages, very promising. Particular emphasis is laid on the study of focus, its functions and manifestations in different subsaharan languages, as well as the differentiation between different types of focus, i.e., term focus (focus on arguments/adjuncts), predicate focus (focus on verb/verb phrase/TAM/truth value), and sentence focus (focus on the whole utterance).

We describe corpora of 25 subsaharan languages created for this purpose, together with ANNIS, the technical infrastructure developed to support linguists in their work with these data collections. ANNIS is specifically designed to support corpora with rich and deep annotation, as IS manifests itself on practically all levels of linguistic description. It provides user-friendly means of querying and visualizations for different kinds of linguistic annotations, including flat, layer-based annotations as used for linguistic glosses, but also hierarchical annotations as used for syntax annotation.

2 Research Activities at the CRC

Within the Collaborative Research Centre, there are several projects eliciting data in large amounts and great diversity. These data, originating from different languages, different modes (written and spoken language) and specific research questions characterize the specification of the linguistic database ANNIS.

2.1 Linguistic Data Base

The project “Linguistic database for information structure: Annotation and Retrieval”, further

database project, coordinates annotation activities in the CRC, provides service to projects in the creation and maintenance of data collections, and conducts theoretical research on multi-level annotations. Its primary goals, however, are the development and investigation of techniques to process, to integrate and to exploit deeply annotated corpora with multiple kinds of annotations. One concrete outcome of these efforts is the linguistic data base ANNIS described further below. For the specific facilities of ANNIS, its application to several corpora of African languages and its use as a general-purpose tool for the publication, visualization and querying of linguistic data, see Sect. 5.

2.2 Gur and Kwa Languages

Gur and Kwa languages, two genetically related West African language groups, are in the focus of the project “Interaction of information structure and grammar in Gur and Kwa languages”, henceforth *Gur-Kwa project*. In a first research stage, the precise means of expression of the pragmatic category *focus* were explored as well as their functions in Gur and Kwa languages. For this purpose, a number of data collections for several languages were created (Sect. 3.1). Findings obtained with this data led to different subquestions which are of special interest from a cross-linguistic and a theoretical point of view. These concern (i) the analysis of syntactically marked focus constructions with features of narrative sentences (Schwarz & Fiedler 2007), (ii) the study of verb-centered focus (i.e., focus on verb/TAM/truth value), for which there are special means of realization in Gur and Kwa (Schwarz, forthcoming), (iii) the identification of systematic *focus-topic-overlap*, i.e., coincidence of focus and topic in sentence-initial nominal constituents (Fiedler, forthcoming). The project's findings on IS are evaluated typologically on 19 selected languages. The questions raised by the project serve the superordinate goal to expand our knowledge of linguistically relevant information structural categories in the less-studied Gur and Kwa languages as well as the interaction between IS, grammar and language type.

2.3 Chadic Languages

The project “Information Structure in the Chadic Languages”, henceforth *Chadic project*, investigates focus phenomena in Chadic

languages. The Chadic languages are a branch of the Afro-Asiatic language family mainly spoken in northern Nigeria, Niger, and Chad. As tone languages, the Chadic languages represent an interesting subject for research into focus because here, intonational/tonal marking – a commonly used means for marking focus in European languages – is in potential conflict with lexical tone, and so, Chadic languages resort to alternative means for marking focus.

The languages investigated in the Chadic project include the western Chadic languages Hausa, Tangale, and Guruntum and the central Chadic languages Bura, South Marghi, and Tera. The main research goals of the Chadic project are a deeper understanding of the following asymmetries: (i) subject focus is obligatorily marked, but marking of object focus is optional; (ii) in Tangale and Hausa there are sentences that are ambiguous between an object-focus interpretation and a predicate-focus interpretation, but in intonation languages like English and German, object focus and predicate focus are always marked differently from each other; (iii) in Hausa, Bole, and Guruntum there is only a tendency to distinguish different types of focus (new-information focus vs. contrastive focus), but in European languages like Hungarian and Finnish, this differentiation is obligatory.

2.4 Focus from a Cross-linguistic Perspective

The project “Focus realization, focus interpretation, and focus use from a cross-linguistic perspective”, further *focus project*, investigates the correspondence between the realization, interpretation and use of with an emphasis on focus in African and south-east Asian languages. It is structured into three fields of research: (i) the relation between differences in realization and differences in semantic meaning or pragmatic function, (ii) realization, interpretation and use of predicate focus, and (iii) association with focus.

The relation between differences in realization and semantic/pragmatic differences (i) particularly pertains the semantic interpretation of focus: For Hungarian and Finnish, a differentiation between two semantic types of foci corresponding to two different types of focus realization was suggested, and we investigate whether the languages studied here have a similar distinction between two (or more) semantic focus types, whether this may differ

from language to language, and whether differences in focus realization correspond to semantic or pragmatic differences.

The investigation of realization, interpretation and use of predicate focus (ii) involves the questions why different forms of predicate focus are often realized in the same way, why they are often not obligatorily marked, and why they are often marked differently from term focus.

Association with focus (iii) means that the interpretation of the sentence is influenced by the focusing of a particular constituent, marked by a focus-sensitive expression (e.g., particles like 'only', or quantificational adverbials like 'always'), while usually, focus does not have an impact on the truth value of a sentence. The project investigates which focus-sensitive expressions there are in the languages studied, what kinds of constituents they associate with, how this association works, and whether it works differently for focus particles and quantificational adverbials.

3 Collections of African Language Data at the CRC

3.1 Gur and Kwa Corpora

The Gur and Kwa corpora currently comprise data from 19 languages.

Due to the scarceness of information available on IS in the Gur and Kwa languages, data had to be elicited, most of which was done during field research, mainly in West Africa, and some in Germany with the help of native speakers of the respective languages. The typologically diverse languages in which we elicited data by ourselves are: Baatonum, Buli, Byali, Dagbani, Ditammari, Gurene, Konkomba, Konni, Nateni, Waama, Yom (Gur languages) and Aja, Akan, Efutu, Ewe, Fon, Foodo, Lelemi, Anii (Kwa languages).

The elicitation of the data based mainly on the questionnaire on information structure, developed by our research group (QUIS, see Section 4.2). This ensured that comparable data for the typological comparison were obtained. Moreover, language-specific additional tasks and questionnaires tailored to a more detailed analysis or language-specific traits were developed.

As the coding of IS varies across different types of texts, different text types were included in the corpus, such as (semi-)spontaneous speech, translations, mono- and dialogues. Most of the languages do not have a long literacy

tradition, so that the corpus data mainly represents oral communication.

In all, the carefully collected heterogeneous data provide a corpus that gives a comprehensive picture of IS, and in particular the focus systems, in these languages.

3.2 Hausar Baka Corpus

In the Chadic project, data from 6 Chadic languages are considered.

One of the larger data sets annotated in the Chadic project is drawn from *Hausar Baka* (Randell, Bature & Schuh 1998), a collection of videotaped Hausa dialogues recording natural interaction in various cultural milieus, involving over fifty individuals of different age and gender. The annotated data set consists of approximately 1500 sentences.

The corpus was annotated according to the guidelines for Linguistic Information Structure Annotation (LISA, see Section 4.2). The Chadic languages show various forms of syntactic displacement, and in order to account for this, an additional annotation level was added: constituents are marked as `ex-situ="+"` if they occur displaced from their canonical, unmarked position.

An evaluation of the focus type and the displacement status reveals tendencies in the morphosyntactic realization of different focus types, see Sect. 5.2.

3.3 Hausa Internet Corpus

Besides these data collections that are currently available in the CRC and in ANNIS, further resources are continuously created. As such, a corpus of written Hausa is created in cooperation with another NLP project of the CRC.

The corpora previously mentioned mostly comprise elicited sentences from little-documented languages with rather small language communities. Hausa, in contrast, is spoken by more than 24 million native speakers, with large amounts of Hausa material (some of it parallel to material in other, more-studied languages) available on the internet. This makes Hausa a promising language for the creation of resources that enable a quantitative study of information structure.

The Hausa internet corpus is designed to cover different kinds of written language, including news articles from international radio stations (e.g., <http://www.dw-world.de>), religious texts, literary prose but also material similar to spontaneous spoken language (e.g., in chat logs).

Parallel sections of the corpus comprise excerpts from the novel Ruwan Bagaja by Abubakar Imam, Bible and Qur'an sections, and the Declaration of Human Rights. As will be described in Section 4.3, these parallel sections open the possibility of semiautomatic morphosyntactic annotation, providing a unique source for the study of information structure in Hausa. Sect. 5.2 gives an example for bootstrapping *ex-situ* constituents in ANNIS only on the basis of morphosyntactic annotation.

4 Data Elicitation and Annotation

4.1 Elicitation with QUIS

The questionnaire on information structure (Skopeteas et al., 2006) provides a tool for the collection of natural linguistic data, both spoken and written, and, secondly, for the elaboration of grammars of IS in genetically diverse languages. Focus information, for instance, is typically elicited by embedding an utterance in a question context. To avoid the influence of a mediator (working) language, the main body of QUIS is built on the basis of pictures and short movies representing a nearly culture- and language-neutral context. Besides highly controlled experimental settings, less controlled settings serve the purpose of eliciting longer, cohesive, natural texts for studying categories such as focus and topic in a near-natural environment.

4.2 Transcription and Manual Annotation

In the CRC, the annotation scheme LISA has been developed with special respect to applicability across typologically different languages (Dipper et al., 2007). It comprises guidelines for the annotation of phonology, morphology, syntax, semantics and IS.

The data mentioned above is, in the case of speech, transcribed according to IPA conventions, otherwise written according to orthographic conventions, and annotated with glosses and IS, a translation of each sentence into English or French, (optionally) additional notes, references to QUIS experiments, and references to audio files and metadata.

4.3 (Semi-)automatic Annotation

As to the automatization of annotation, we pursue two strategies: (i) the training of classifiers on annotated data, and (ii) the projection of annotations on texts in a source language to parallel texts in a target language.

Machine Learning. ANNIS allows to export query matches and all their annotated features to the table format ARFF which serves as input to the data mining tool WEKA (Witten & Frank, 2005), where instances can be clustered, or used to train classifiers for any annotation level.

Projection. Based on (paragraph-, sentence- or verse-) aligned sections in the Hausa internet corpus, we are about to project annotations from linguistically annotated English texts to Hausa, in a first step parts of speech and possibly nominal chunks. On the projected annotation, we will train a tagger/chunker to annotate the remaining, non-parallel sections of the Hausa internet corpus. Existing manual annotations (e.g. of the Hausar Baka corpus) will then serve as a gold standard for evaluation purposes.

Concerning projection techniques, we expect to face a number of problems: (i) the question how to assign part of speech tags to categories existing only in the target language (e.g., the person-aspect complex in Hausa that binds together information about both the verb (aspect) and its (pronominal subject) argument), (ii) issues of orthography: the official orthography Hausa (*Boko*) is systematically underspecified wrt. linguistically relevant distinctions. Neither vowel length nor different qualities of certain consonants (*r*) are represented, and also, there is no marking of tones (see Examples 1 and 2, fully specified word forms in brackets). To distinguish such homographs, however, is essential to the appropriate interpretation and linguistic analysis of utterances.

(1) **ciki** - 1. [cíkii, noun]
stomach, 2. [cíkí, prep.]
inside

(2) **dace** - 1. [dàacée, noun]
coincidence, 2. [dáacèè, verb]
be appropriate

We expect that in these cases, statistical techniques using context features may help to predict correct vowelization and tonal patterns.

5 ANNIS – the Linguistic Database of Information Structure Annotation

5.1 Conception and Architecture

ANNIS (ANNotation of Information Structure) is a web-based corpus interface built to query and visualize multilevel corpora. It allows the user to formulate queries on arbitrary, possibly nested annotation levels, which may be

conflictingly overlapping or discontinuous. The types of annotations handled by ANNIS include, among others, flat, layer-based annotations (e.g., for glossing) and hierarchical trees (e.g., syntax).

Source data. As an architecture designed to facilitate diverse and integrative research on IS, ANNIS can import formats from a broad variety of tools from NLP and manual annotation, the latter including EXMARaLDA (Schmidt, 2004), annotate (Brants and Plaehn, 2000), Synpathy (www.lat-mpi.eu/tools/synpathy/), MMAX2 (Müller and Strube, 2006), RSTTool (O'Donnell, 2000), PALinkA (Orasan, 2003), Toolbox (Busemann & Busemann, 2008) etc. These tools allow researchers to annotate data for syntax, semantics, morphology, prosody, phonetics, referentiality, lexis and much more, as their research questions require.

All annotated data are merged together via a **general interchange format** PAULA (Dipper 2005, Dipper & Götze 2005), a highly expressive standoff XML format that specifically allows further annotation levels to be added at a later time without disrupting the structure of existing annotations. PAULA, then, is the native format of ANNIS.

Backend. The ANNIS server uses a relational database that offers many advantages including full Unicode support and regular expression searches. Extensive search functionalities are supported, allowing complex relations between individual word forms and annotations, such as all forms of overlapping, contained or adjacent annotation spans, dominance axes (children, ancestors etc., as well as common parent, left- or right-most child and more), etc.

Search. In the user interface, queries can be formulated using the ANNIS Query Language (AQL). It is based on the definition of nodes to be searched for and the relationships between these nodes (see below for some examples). A graphical query builder is also included in the web interface to make access as easy as possible.

Visualization. The web interface, realized as a window-based AJAX application written in Java, provides visualization facilities for search results. Available visualizations include token-based annotations, layered annotations, tree-like annotations (directed acyclic graphs), and a discourse view of entire texts for, e.g., coreference annotation. Multimodal data is represented using an embedded media player.

Special features. By allowing queries on multiple, conflicting annotation levels simultaneously, the system supports the study of interdependencies between a potentially limitless variety of annotation levels.

At the same time, ANNIS allows us to integrate and to search through heterogeneous resources by means of a unified interface, a powerful query language and a intuitive graphical query editor and is therefore particularly well-suited for the purpose of language documentation. In particular, ANNIS can serve as a tool for the publication of data collections via internet. A fine-grained user management allows granting privileged users access to specific data collections, to make a corpus available to the public, or to seal (but preserve) a resource, e.g., until legal issues (copyright) are settled. This also makes publishing linguistic data collections possible without giving them away.

Moreover, ANNIS supports deep links to corpora and corpus queries. This means that queries and query results referred to in, e.g., a scientific paper, can be reproduced and quoted by means of (complex) links (see following example).

5.2 Using ANNIS. An Example Query

As an illustration for the application of ANNIS to the data collections presented above, consider a research question previously discussed in the study of object focus in Hausa.

In Hausa, object focus can be realized in two ways: either *ex-situ* or *in-situ* (Section 3.2). It was found that these realizations do not differ in their semantic type (Green & Jaggard 2003, Hartmann & Zimmermann 2007): instead, the marked form signals that the focused constituent (or the whole speech act) is unexpected for the hearer (Zimmermann 2008). These assumptions are consistent with findings for other African languages (Fiedler et al. 2006).

In order to verify such claims on corpora with morphosyntactic and syntactic annotation for the example of Hausa, a corpus query can be designed on the basis of the Hausar Baka corpus that comprises not only annotations for grammatical functions and information-structural categories, but also an annotation of *ex-situ* elements.

Ee . Ee wàllaahi . Dooguuwar riigaa nakèe sòò dà d'an kwaalii .

exmaralda										
Select Displayed Annotation Levels ▾										
CLASS	PTC	PTC		A	N	PRONPRS	V	P	N	
EX-SITU				+						
FOCUS	cf-conf			nf-sol					nf-sol	
GIVEN	new			new		acc-sit	new		new	
GLOSS	yes	yes	by.God	long.F-of	gown	1.SG-PROG.REL	want	with	scarf	
MORPH	ee	ee	wallaahi	dooguuwa-r	riigaa	na-kee	soo	dà	d'an	kwaalii
TOPE	H	H	LHL	HHH	LH	HL	HL	L	H	HH
TRANSLATION	Yes.	Yes, that's true.		I'd like a long dress and a (matching) scarf.						
tok	Ee	Ee	wàllaahi	Dooguuwar	riigaa	nakèe	sòò	dà	d'an	kwaalii

Figure 1: ANNIS partitur view, Hausar Baka corpus.

So, in (3), we look for ex-situ constituents (variable #1) in declarative sentences in the Hausa Bakar corpus, i.e., sentences that are not translated as questions (variable #2) such that #1 is included in #2 (#1 *_i_* #2).

```
(3) EX-SITU="+" &
TRANSLATION=".*[?]" & #1
_i_ #2
```

Considering the first 25 matches for this query on Hausar Baka, 16 examples reveal to be relevant (excluding interrogative pronouns and elliptical utterances). All of these are directly preceded by a period (sentence-initial) or a comma (preceded by *ee* ‘yes’, interjections or exclamations), with one exception, preceded by a sentence initial negation marker.

Only seven examples are morphologically marked by focus particles (*nee*, *cee*), focus-sensitive adverbs (*kawàì* ‘only’) or quantifiers (*koomee* ‘every’). In nine cases, a personal pronoun follows the ex-situ constituent, followed by the verb. Together, these constraints describe all examples retrieved, and as a generalization, we can now postulate a number of patterns that only make use of morphosyntactic and syntactic annotation (token *tok*, morphological segmentation *MORPH*, parts of speech *CLASS*, nominal chunks *CHUNK*) with two examples given below:

```
(4) tok=/[ , . ! ? ] / &
CHUNK="NC" & MORPH=/[cn]ee/ &
#1 . #2 & #2 . #3
(5) tok=/[ , . ! ? ] / &
CHUNK="NC" & CLASS=/PRON.* / &
CLASS=/V / & #1 . #2 & #2 .
#3 & #3 . #4
```

In (4), we search for a nominal chunk following a punctuation sign and preceding a

focus particle (*cee* or *nee*), in (5), we search for a nominal chunk preceding a sequence of pronoun/aspect marker and verb.

One example matching template (5) from the Hausar Baka corpus is given in Fig. 1.

While AQL can be used in this way to help linguists understand the grammatical realization of certain phenomena, and the grammatical context they occur in, patterns like (5) above are probably not too readable to an interested user. This deficit, however, is compensated by the graphical query builder that allows users to create AQL queries in a more intuitive way, cf. Fig. 2.

Of course, these patterns are not exhaustive and overgenerate. However, they can be directly evaluated against the manual ex-situ annotation in the Hausar Baka corpus and further refined.

So, the manual annotation of ex-situ constituents in the Hausar Baka corpus provides templates for the semi-automatic detection of ex-situ constituents in a morphosyntactically annotated corpus of Hausa: The patterns generate a set of candidate examples from which a human annotator can then choose real ex-situ constituents. Indeed, for a better understanding of ex-situ object focus, a study with a larger database of more natural language would be of great advantage, and this pattern-based approach represents a way to create such a database of ex-situ constructions in Hausa.

Finally, it would also help find instances of predicate focus. When a V(P) constituent is focused in Hausa, it is nominalized, and fronted like a focused nominal constituent (Hartmann & Zimmermann 2007).

5.3 Related Corpus Tools

Some annotation tools come with search facilities, e.g. **Toolbox** (Busemann & Busemann, 2008), a system for annotating, managing and

analyzing language data, mainly geared to lexicographic use, and ELAN (Hellwig et al., 2008), an annotation tool for audio and video data.

In contrast, ANNIS is not intended to provide annotation functionality. The main reason behind this is that both Toolbox and ELAN are problem-specific annotation tools with limited capabilities for application to different phenomena than they were designed for. Toolbox provides an intuitive annotation environment and search facilities for flat, word-oriented annotations; ELAN, on the other hand, for annotations that stand in a temporal relation to each other.

These tools – as well as the other annotation tools used within the CRC – are tailored to a particular type of annotation, neither of them being capable of sufficiently representing the data from all other tools. Annotation on different levels, however, is crucial for the investigation of information structural phenomena. In order to fill in this gap, ANNIS was designed primarily with the focus on visualization and querying of multi-layer annotations. In particular, ANNIS allows to integrate annotations originating from *different tools* (e.g., syntax annotation created with Synpathy, coreference annotation created with MMAX2, and flat, time-aligned annotations created with ELAN) that nevertheless refer to the *same primary data*. In this respect, ANNIS, together with the data format PAULA and the libraries created for the work with both, is best compared to general annotation frameworks such as ATLAS, NITE and LAF.

Taking the **NITE XML Toolkit** as a representative example for this kind of frameworks, it provides an abstract data model, XML-based formats for data storage and metadata, a query language, and a library with JAVA routines for data storage and manipulation, querying and visualization. Additionally, a set of command line tools and simple interfaces for corpus querying and browsing are provided, which illustrates how the libraries can be used to create one's own, project-specific corpus interfaces and tools.

Similarly to ANNIS, NXT supports time-aligned, hierarchical and pointer-based annotation, conflicting hierarchies and the embedding of multi-modal primary data. The data storage format is based on the bundling of multiple XML files similar to the standoff concept employed in LAF and PAULA.

One fundamental difference between NXT and ANNIS, however, is to be seen in the primary

clientele it targets: The NITE XML Toolkit is aimed at the developer and allows to build more specialized displays, interfaces, and analyses as required by their respective end users when working with highly structured data annotated on multiple levels.

As compared to this, ANNIS is directly targeted at the end user, that is, a linguist trying to explore and to work with a particular set of corpora. Therefore, an important aspect of the ANNIS implementation is the integration with a data base and convenient means for visualization and querying.

6 Conclusion

In this paper, we described the Africanist projects of the CRC „Information Structure“ at the University of Potsdam and the Humboldt University of Berlin/Germany, together with their data collections from currently 25 subsaharan languages. Also, we have presented the linguistic database ANNIS that can be used to publish, access, query and visualize these data collections. As one specific example of our work, we have described the design and ongoing construction of a corpus of written Hausa, the Hausa internet corpus, sketched the relevant NLP techniques for (semi)automatic morphosyntactic annotation, and the application of the ANNIS Query Language to filter out ex-situ constituents and their contexts, which are relevant with regard to our goal, a better understanding of focus and information structure in Hausa and other African languages.

References

- T. Brants, O. Plaehn. 2000. Interactive Corpus Annotation. In: *Proc. of LREC 2000*, Athens, Greece.
- A. Busemann, K. Busemann. 2008. Toolbox Self-Training. Technical Report (Version 1.5.4 Oct 2008) <http://www.sil.org/>
- J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, H. Voormann. 2003. The NITE XML Toolkit: Flexible Annotation for Multi-modal Language Data. *Behavior Research Methods, Instruments, and Computers* 35(3), 353-363.
- J. Carletta, S. Evert, U. Heid, J. Kilgour. 2005. The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal*, 313-334.

- S. Dipper. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: *Proc. of Berliner XML Tage 2005 (BXML 2005)*, Berlin, Germany, 39-50.
- S. Dipper, M. Götze. 2005. Accessing Heterogeneous Linguistic Data - Generic XML-based Representation and Flexible Visualization. In *Proc. of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznan, Poland, 206-210.
- S. Dipper, M. Götze, S. Skopeteas (eds.). 2007. *Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Interdisciplinary Studies on Information Structure 7, 147-187. Potsdam: University of Potsdam.
- I. Fiedler. forthcoming. Contrastive topic marking in Gbe. In *Proc. of the 18th International Conference on Linguistics, Seoul, Korea, 21. - 26. July 2008*.
- I. Fiedler, K. Hartmann, B. Reineke, A. Schwarz, M. Zimmermann.. forthcoming. Subject Focus in West African Languages. In M. Zimmermann, C. Féry (eds.), *Information Structure. Theoretical, Typological, and Experimental Perspectives*. Oxford: Oxford University Press. .
- M. Green, P. Jaggar. 2003. Ex-situ and in-situ focus in Hausa: syntax, semantics and discourse. In J. Lecarme et al. (eds.), *Research in Afroasiatic Grammar 2 (Current Issues in Linguistic Theory)*. Amsterdam: John Benjamins. 187-213.
- K. Hartmann, M. Zimmermann. 2004. Nominal and Verbal Focus in the Chadic Languages. In F. Perrill et al. (eds.), *Proc. of the Chicago Linguistics Society*. 87-101.
- K. Hartmann, M. Zimmermann. 2007. In Place - Out of Place? Focus in Hausa. In K. Schwabe, S. Winkler (eds.), *On Information Structure, Meaning and Form: Generalizing Across Languages*. Benjamins, Amsterdam: 365-403.
- B. Hellwig, D. Van Uytvanck, M. Hulsbosch. 2008. ELAN – Linguistic Annotator. Technical Report (as of 2008-07-31). <http://www.lat-mpi.eu/tools/elan/>
- É. Kiss. 1998. Identificational Focus Versus Information Focus. *Language* 74: 245-273.
- M. Krifka. 1992. A compositional semantics for multiple focus constructions, in Jacobs, J: *Informationsstruktur und Grammatik*, Opladen, Westdeutscher Verlag, 17-53.
- C. Müller, M. Strube. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In: S. Braun et al. (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 197–214.
- M. O'Donnell. 2000. RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory. In: *Proc. of the International Natural Language Generation Conference (INLG'2000)*, 13-16 June 2000, Mitzpe Ramon, Israel, 253–256.
- C. Orasan. 2003. Palinka: A Highly Customisable Tool for Discourse Annotation. In: *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan.
- R. Randell, A. Bature, R. Schuh. 1998. Hausar Baka. <http://www.humnet.ucla.edu/humnet/aflang/hausarbaka/>
- T. Schmidt. 2004. Transcribing and Annotating Spoken Language with Exmaralda. In: *Proc. of the LREC-Workshop on XML Based Richly Annotated Corpora, Lisbon 2004*. Paris: ELRA.
- A. Schwarz. Verb and Predication Focus Markers in Gur. forthcoming. In I. Fiedler, A. Schwarz (eds.), *Information structure in African languages (Typological Studies in Language)*, 307-333. Amsterdam, Philadelphia: John Benjamins.
- A. Schwarz, I. Fiedler. 2007. Narrative focus strategies in Gur and Kwa. In E. Aboh et al. (eds.): *Focus strategies in Niger-Congo and Afroasiatic – On the interaction of focus and grammar in some African languages*, 267-286. Berlin: Mouton de Gruyter.
- S. Skopeteas, I. Fiedler, S. Hellmuth, A. Schwarz, R. Stoel, G. Fanselow, C. Féry, M. Krifka. 2006. *Questionnaire on information structure (QUIS)*. Interdisciplinary Studies on Information Structure 4. Potsdam: University of Potsdam.
- I. H. Witten, E. Frank, *Data mining: Practical machine learning tools and techniques*, 2nd edn, Morgan Kaufman, San Francisco, 2005.
- M. Zimmermann. 2008. Contrastive Focus and Emphasis. In *Acta Linguistica Hungarica* 55: 347-360.